

半监督学习——提出

传统机器学习分两类：监督学习、无监督学习

监督学习：充分的标记样本才能保证其训练精度

无监督学习：仅仅是对未标记样学习，不能保证精度

现实：有大量的无标签样本和少量的标记样本

“廉价的”未标记样本是否是无用的？

否，是对数据资源是一种浪费。

因此如何有效的同时利用两种样本进行学习被研究者的关注——半监督学习

半监督学习——提出

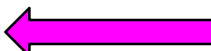
二十世纪九十年代，半监督学习（**Semi-supervised Learning**）被提出。

- 半监督学习研究主要关注当训练数据的部分信息缺失（包括数据的类别标签缺失、数据的部分特征维缺失、噪声等）的情况下，如何获得具有良好性能和泛化能力的学习机器，即利用大量的未标记样本来辅助标记样本来建立一个好的学习器。
- 应用领域：网页检索和文本分类、基于生物特征的身份识别、医学数据处理、数字图像处理、视频标签等

半监督学习——定义

- 半监督学习的基本思想是利用数立学习器对未来标签样本进行标
- 形式化描述为：

给定一个来自某未知分布的样
其中 L 是已标签样本集 $L=\{(x$
 U 是一个未标签样本集 $U=\{x'$

$L \rightarrow L_1 \rightarrow L_2 \rightarrow \dots \rightarrow L_n$  $U \rightarrow U_1 \rightarrow U_2 \rightarrow \dots \rightarrow U_n$

希望得到函数 $f: X \rightarrow Y$ 可以准确的对样本 x 预测标签 y

当前标记不足，难以建立满意的分类器

策略：逐步增加标记数据，最终得到满意的分类器

因此，**如何获得正确的标记数据是关键**



半监督学习——假设

- 半监督学习问题从样本的角度而言是利用少量标注样本和大量未标注样本进行机器学习，从概率学习角度可理解为研究如何利用训练样本的输入边缘概率 $P(\mathbf{x})$ 和条件输出概率 $P(\mathbf{y} | \mathbf{x})$ 的联系设计具有良好性能的分类器。

这种联系的存在是建立在某些假设的基础上

即聚类假设 (cluster assumption)

平滑假设 (Smoothness Assumption)

流形假设 (manifold assumption)

6.5 半监督学习——平滑假设

平滑假设：位于稠密数据区域的两个距离很近的样例的类标签相似，也就是说，
当两个样例被稠密数据区域中的边连接时，它们大概率具有相同的类标签；
相反地，当两个样例被稀疏数据区域分开时，它们的类标签趋于不同。

这一假设反映了决策函数的局部平滑性。和聚类假设着眼整体特性不同，平滑假设主要考虑模型的局部特性。

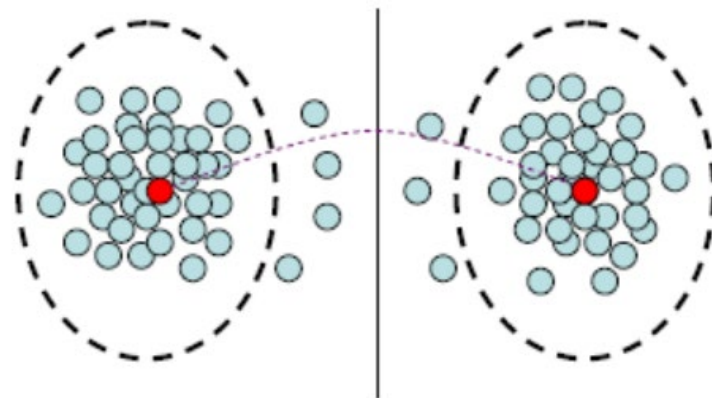
在该假设下，大量未标记示例的作用是让数据空间变得更加稠密，从而有助于更加准确地刻画局部区域的特性
使得决策函数能够更好地进行数据拟合

6.5 半监督学习——聚类假设

- 聚类假设（平滑假设特例）：是指处在**相同聚类中的样本有较大的可能拥有相同的标记**

根据该假设，决策边界就应该尽量通过数据较为稀疏的地方，从而避免把稠密的聚类中的数据点分到决策边界两侧。

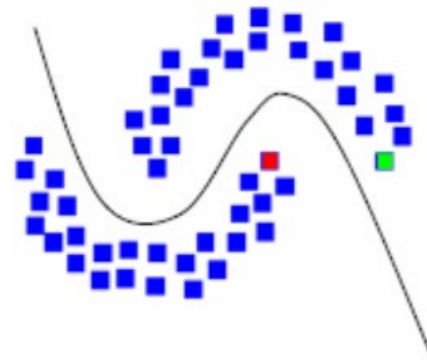
因此，大量未标记样本的作用就是帮助探明样本空间中数据分布的稠密和稀疏区域



6.5 半监督学习——流行假设

- 流形假设：输入空间由多个低维流形组成，位于同一流形上的数据点具有相同标签。
- 将高维数据嵌入到低维流形中，当两个样例位于低维流形中的一个小局部邻域内时，它们具有相似的类标签。

这一假设与平滑性假设相似



半监督学习——常用学习算法

- 半监督学习算法按照不同的模型假设,可以大致将现有的半监督学习算法分为五类:
 - EM
 - 自学习(Self-training)
 - 基于生成模型的方法(EM with generative mixture models)
 - 协同训练(Co-training)
 - 直推式支持向量机 (Transductive Support Vector Machines)
 - 基于图的方法

半监督学习——自学习算法

- 自学习要表达的核心思想是在分类器递归拟合的时候，每次递归仅将满足设定的置信度阈值的即置信度高的样本纳入到已标记样本集中，参与递归拟合。
- 算法流程：
 - Step1:用已标记的样本来训练得到一个初始分类器；
 - Step2:用初始分类器对未标记样本进行分类，将标记**置信度高**的未标记样本进行标记；
 - Step3:对所有样本进行重新训练，直到将所有未标记样本

半监督学习——自学习算法

- 自学习优点：

1. 最简单的半监督学习方法，效果不错。
2. 这是一种wrapper方法，可以应用到已有的（复杂）分类器上。

- 自学习缺点：

1. 早期的错误会强化——>启发式的缓解方案：如果数据的置信分数低于某个阈值再把它的标签去掉。
2. 在收敛性方面没有保障。——>但是也有特例，自我训练等价于EM算法。有部分存在封闭解的特殊情况。

半监督学习——协同训练算法（多视角）

此类算法隐含地利用了聚类假设或流形假设
使用两个或多个学习器，在学习过程中，这些学习器
挑选若干置信度高的未标记示例进行互相标记，从而
更新模型

最早提出**Co-training**的是A. Blum和T. Mitchell
南大周志华提出了**Tri-training**

该算法的一个显著特点是使用了三个分类器，不仅可以
简便地处理标记置信度估计问题以及对未见示例的
预测问题，还可以利用集成学习Ensemble learning
来提高学习的泛化能力

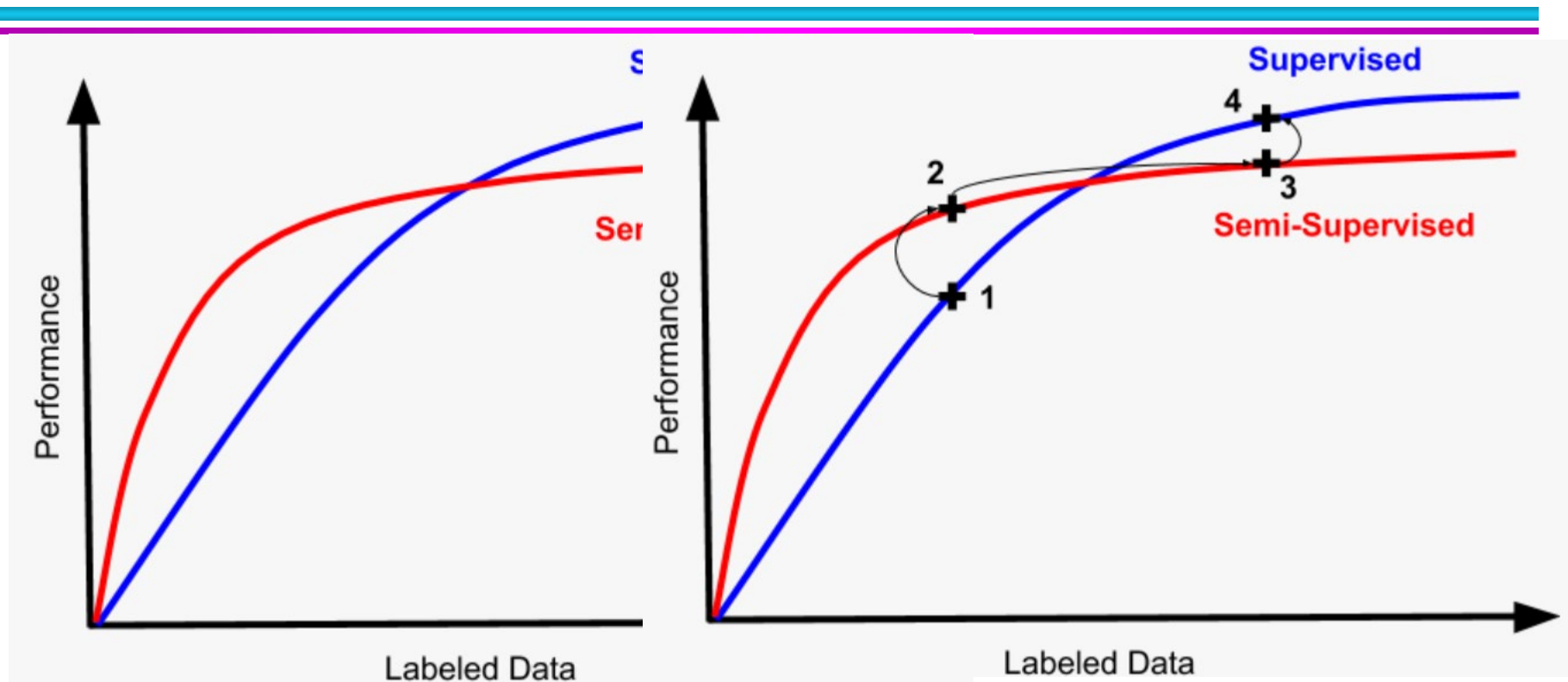
半监督学习——协同训练算法

- 算法流程：
- 步骤一：对标记样本进行可重复取样以获得三个有标记样本集，从每个样本集训练出一个分类器。
- 步骤二：在协同训练过程中，各分类器所获得的新标记示例都由其余两个分类器协作提供，具体来说，如果两个分类器对同一个未标记示例的预测相同，则该示例就被认为具有较高的标记置信度，并在标记后被加入第三个分类器的有标记训练集。以便对方利用这些新标记的示例进行更新。

6.5 半监督学习——Tri-training算法

- 算法流程：
- 步骤一：对标记样本进行可重复取样以**获得三个**有标记样本集，从每个样本集训练出一个分类器。
- 步骤二：在协同训练过程中，各分类器所获得的新标记示例都由其余两个分类器协作提供
具体来说：
若两个分类器对同一个未标记示例的**预测相同**，则该示例就被认为**具有较高的标记置信度**，并在标记后被**加入第三个分类器**的有标记训练集。以便对方利用这些新标记的示例进行更新。

6.5 半监督学习——面临的问题



- 标注数据不多时，**SS**可以带来一定的性能提升。
- 但提升只能帮你把模型表现从「不可接受」提高到「稍微好了那么一点、但还是没办法使用」
- 使用了**ss**面对更多的标注数据时，性能增长曲线比监督学习更平缓；原因之一是无标注数据可能会带来偏倚

如何理解有监督、无监督、半监督、弱监督学习

正常使用主观题需2.0以上版本雨课堂